

Man Made Language Models? Evaluating LLMs’ Perpetuation of Masculine Generics Bias

Enzo Doyen and Amalia Todirascu

LiLPa, University of Strasbourg
{enzo.doyen,todiras}@unistra.fr

Abstract

Instruct-based large language models (LLMs) have been shown to propagate and even amplify gender bias when prompted with contextually constrained instructions (e.g., writing a text from a description or selecting a gendered pronoun). However, little attention has been paid to biases in responses to contextually unconstrained (generic) instructions conveyed by gendered language, particularly masculine generics (MG). MG, found in many gender-marked languages, denote the use of the masculine gender as a supposedly neutral reference to mixed-gender groups or individuals whose gender is unknown or non-binary. Yet, psycholinguistic studies demonstrate that MG are not neutral and systematically induce gender bias. This study investigates how both local and proprietary LLMs are MG-biased when responding to generic prompts in French, examining LLMs’ MG bias rates and use of gender-fair language (GFL). We create a 16k+ human noun database from existing lexical resources and evaluate six LLMs on four instruction-response datasets under two conditions: prompts with and without MG. Overall, we find that $\approx 27.57\%$ of LLMs’ responses to MG-filtered generic instructions are MG-biased ($\approx 78.55\%$ with MG-containing prompts). Moreover, we find that LLMs rarely use GFL spontaneously. These findings highlight the persistence of MG bias in LLM outputs and models’ limited tendency towards GFL strategies.

 [spidersouris/llm_masc_gen](https://github.com/spidersouris/llm_masc_gen)

1 Introduction

Masculine generics (MG) are a linguistic feature found in many gender-marked languages, among which French, German, or Dutch. MG denote the use of the masculine gender in gendered languages as a “default” or supposedly neutral gender to refer to either a) a mixed group of men and women, or b) a person whose gender is irrelevant, non-binary

or unknown within the context, as in the following examples in French:

- (a) **Les étudiants**^{MG} ont participé à l’atelier.
(**Students**^{MG} participated to the workshop.)
- (b) **Un athlète**^{MG} doit s’entraîner régulièrement pour progresser.
(**An athlete**^{MG} needs to train regularly to progress.)

Comparable equivalents of MG in English would translate into c) consistently using occupation nouns ending with the “-man” suffix or, albeit old-fashioned, d) the generic “he” instead of “they”:

- (c) **The policemen**^{MG} started the training.
- (d) The reader may skip this chapter if **he**^{MG} wishes.

While such usage of the masculine gender supposedly acts as a neutralizer and contrasts with masculine specifics (i.e., noun phrases referring specifically to men)¹, empirical research in psycholinguistics shows that this supposed genericness is not processed as such by native speakers (Gygax et al., 2008; Glim et al., 2024; Rothermund and Strack, 2024), precisely due to the generic and the specific forms being morphologically (but not semantically) identical (Irmen and Kurovskaja, 2010; Gygax et al., 2021). Psycholinguistics studies prove that MG induce cognitive biases that amplify male-centric mental representations at the expense of women (Braun et al., 2005; Gygax et al., 2012; Tibblin et al., 2023).

The rapid development of large language models (LLMs) and their remarkable capabilities across various textual tasks have made them indispensable tools for numerous applications. Relatedly, the release of ChatGPT by OpenAI and similar chatbots through user-friendly web applications has made it extremely trivial for anyone to generate textual content. However, these models inherently reflect biases present in their training data, including those

¹For example, in the sentence “**Cet étudiant** a participé à l’atelier.” (“**This student** participated to the workshop.”), the use of the masculine demonstrative determiner “cet” specifically indicates that the student is male.

related to gender representation. Their widespread adoption introduces at least three distinct pathways through which such biases may propagate:

(1) Exposure to AI-generated content. As a growing amount of text on the Internet is LLM-generated (Brooks et al., 2024; Wei and Tyson, 2024; Sun et al., 2025; Dolezal et al., 2026), people consuming this content are increasingly exposed to the related biases.

(2) LLM training on synthetic data. The use of synthetically generated data in the training of LLMs (Long et al., 2024; Nadas et al., 2025) raises the risk of recursively reinforcing existing model biases.

(3) Direct interaction with conversational LLMs. Many users engage with LLMs primarily for informational and decision-support purposes in conversational contexts, with Chatterji et al. (2025) reporting that 73% of usage is non-work-related and largely interaction-based. Such chat-like exchanges may contain biased responses and could influence users’ perceptions, judgments, and decisions in real time.

While previous research in Natural Language Processing (NLP; see Section 2) largely discussed the issues surrounding the propagation and amplification of gender biases by LLMs in constrained contexts (e.g., instructing an LLM to continue a gender-ambiguous sentence, or to translate from/to languages with different gender systems), a crucial gap exists in the literature when it comes to the biases conveyed by LLMs’ responses to contextually unconstrained (or generic) instructions due to MG use. We define generic instructions as prompts a) devoid of any reference to specific individuals and b) framed around common tasks or questions (“Practical Guidance and Seeking Information” as per Chatterji et al. (2025))².

We focus on only one language with MG (French), but our methodology can be extended to any other gender-marked language given appropriate resources, that is a human noun dataset (presented in Section 3.1), and filtered instruction/output pairs datasets (presented in Section 3.2). Using these resources, we query six LLMs with filtered MG or non-MG instructions, detect MG and gender-fair language (GFL) forms (inclusive and neutral) in their responses, and validate human noun usage in context. We then evaluate the use

of MG and GFL by LLMs across two experiments, drawing a comparison with human-written datasets using a novel Pareto analysis evaluation framework to take into account the multiple complementary dimensions of GFL.

2 Related Work

MG and the associated male bias have been extensively studied in the field of psycholinguistics for a variety of gender-marked languages (Silveira, 1980; Stahlberg et al., 2001; Safina, 2024), including French (Gygax et al., 2012; Richy and Burnett, 2021; Xiao et al., 2023). For example, Gygax et al. (2008) tested English, French, and German speakers with role nouns (e.g., “social workers”; MG in French and German). When later asked to judge continuations referring to group members (e.g., “women”), English participants’ interpretations followed stereotypes (e.g., “beauticians” → female, “politicians” → male), while French and German participants were strongly influenced by grammatical gender: masculine plural forms consistently produced male-biased interpretations, even for female-stereotyped roles. Similarly, a survey conducted by Lévy et al. (2017) on the use of GFL in French showed a positive effect on mental representations. Respondents asked to name a celebrity after inclusive or gender-neutral formulations (e.g., both gendered forms, or non-gendered nouns) more often named women, whereas MG formulations prompted male-centric answers. Similar findings were reported by Stahlberg et al. (2001) for German using MG role nouns.

In NLP, gender bias is by far the most studied type of bias (Ducel et al., 2024b). Studies have drawn attention to gender bias in word embeddings (Bolukbasi et al., 2016), in machine translation systems (Wisniewski et al., 2021; Vanmassenhove, 2024; Savoldi et al., 2025) and in text classification tasks (Sobhani and Delany, 2024). More recently, LLMs too have been shown to exhibit gender biases and convey stereotypes when generating context-restricted textual content (see Kotek et al. (2023) for pronoun disambiguation; Döll et al. (2024) for pronoun prediction; You et al. (2024) for neutral name prediction), including in languages other than English (Ducel et al., 2024a; Ding et al., 2025). For example, in German, Schmidt and Hess (2026) have proposed an LLM-as-a-judge framework for measuring gender neutrality in LLM outputs, and reported that a restrictive gender-neutrality system

²Examples of such instructions drawn from datasets used in this study are presented in Appendix E.

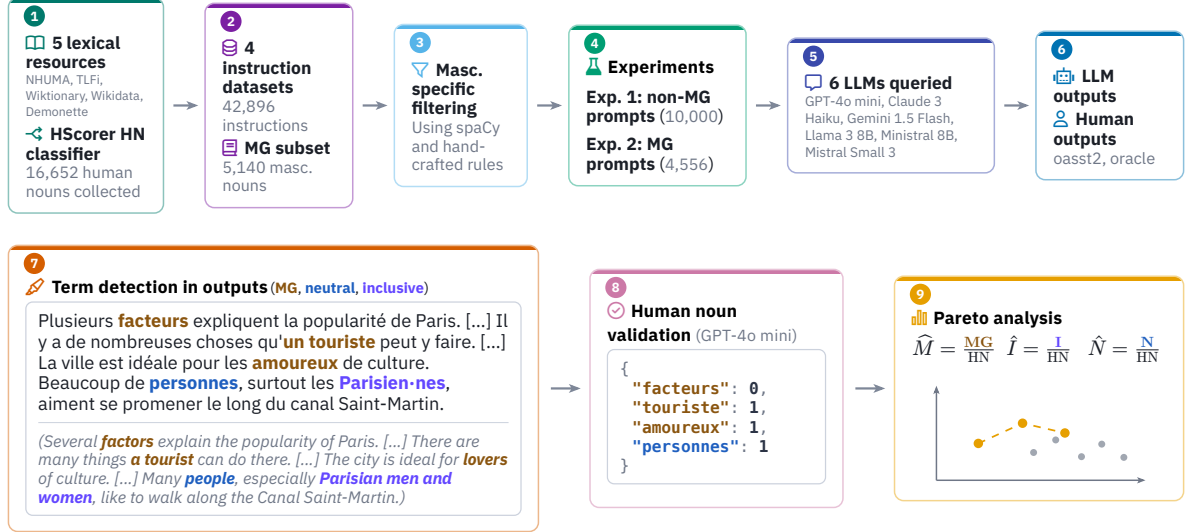


Figure 1: Overview of our methodological pipeline. We create a HN database from existing lexical resources and a custom classifier, and use a subset comprised of masculine nouns to filter out instructions referring to specific male individuals. We then divide generic instructions into two groups based on whether they contain MG or not, and query 6 LLMs. We automatically analyze LLM and human outputs for MG and GFL usage, validate human nouns with GPT-4o mini to resolve polysemic ambiguities, and compare which models and datasets are the most gender-fair using a Pareto analysis.

prompt substantially increased neutral and mixed forms in LLM responses compared to a standard system prompt. Nonetheless, no studies have so far quantified MG and GFL use by LLMs in responses to generic user instructions, nor examined to what extent LLMs are prone to propagating MG-related gender bias in conversational contexts, a gap this work aims to bridge.

3 Methodology

Our methodology is detailed in the sections below. In Section 3.1, we create a database of French human nouns (HN) using existing online French lexical resources, and we build a French HN classifier, HScorer, to filter out nouns not referring to human beings. In Section 3.2, we present LLM-generated and human-written instruction/output datasets used for our experiments and we detail our filtering process to remove masculine specific occurrences (that is, contexts where the masculine gender refers to specific male individuals and is not generic). In Section 3.3, we introduce our pipeline to query LLMs with the previously filtered output datasets, automatically detect MG and inclusive terms in both LLMs’ responses and human-written responses for comparison, and validate HNs in these responses with GPT-4o mini. Finally, in Section 3.4, we show how we measure MG bias in

human-written and LLM-inferred outputs using a novel Pareto analysis evaluation design. The complete pipeline is presented in Figure 1.

3.1 Database of French Human Nouns and HScorer

Since MG are nouns used strictly to refer to human beings, the first step was to create a database of French HNs to leverage for detection. To our knowledge, the only existing database in French is that of the NHUMA project (Stosic and Lagae, 2013)³, which aims to provide an extensive linguistic description of French HNs. However, this database has a limited number of entries (1,107), many of which are not MG. We used it as a basis and extended it with data scraped or collected from French lexical databases: TLFi⁴, a digitalized version of the *Trésor de la langue française* dictionary; French Wiktionary⁵; and French Wikidata⁶ (details in Appendix A).

In addition to these resources, and to further extend our dictionary, we also leveraged the Demonette database (Namer et al., 2023), which contains pairs of masculine/feminine nouns. We also performed recursive search in TLFi to get nouns whose

³<https://nomsdhumains.weebly.com/>

⁴<http://atilf.atilf.fr/>

⁵<https://fr.wiktionary.org/>

⁶<https://wikidata.org/>

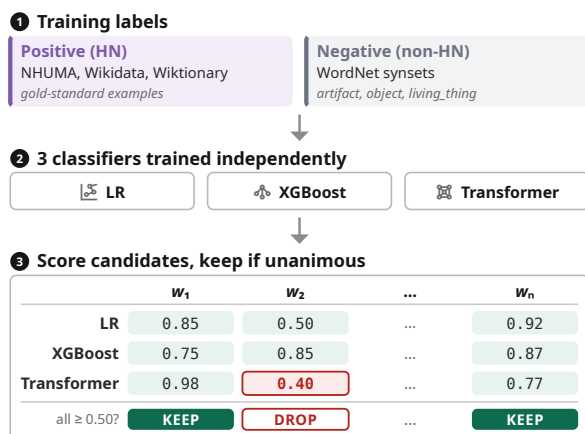


Figure 2: Overview of HScorer. We use golden HN (NHUMA, Wikidata, Wiktionary) and non-HN (WordNet) datasets to train an LR, an XGBoost and a Transformer CamemBERT model. Only words w_n with full agreement as HN (≥ 0.5) from all three classifiers are added to the database.

definition started with one of the previously TLFi-scraped nouns. However, the Demonette database includes pairs of words that do not necessarily refer to human beings (animal nouns, mythology-related nouns). Similarly, the recursive search performed on TLFi led to a number of false positives.⁷ As a potential solution to these false positives, we built an HN classifier to detect whether a noun is commonly used to refer to a human being, **HScorer**. Figure 2 presents the classifier.

HScorer relies on three types of models: a logistic regression (LR) model, an XGBoost model (Chen and Guestrin, 2016) and a Transformer model (Vaswani et al., 2017). The LR and XGBoost model types were chosen based on their best performance compared to a set of classic machine learning model types (see Table 6 in Appendix for cross-validation results). For the Transformer model, we used CamemBERT (Martin et al., 2020) for its best performance with French. HScorer aims at classifying nouns from the Demonette (8,048 entries)⁸ and TLFi Recursive (13,961 entries) datasets as human or non-human. Only words classified as HNs by the three models (full agreement) would be part of the final HN database.

Each model was trained on golden HN and non-HN datasets. The golden HN dataset was built

from previously scraped entries from Wikidata and Wiktionary, and entries from the NHUMA database. For the golden non-HN dataset, we leveraged the WordNet database (Princeton University, 2010) and its collection of universal synsets through NLTK (Bird and Loper, 2004) to filter out French human-related nouns and retrieve non-HN with synsets `artifact.n.01`, `object.n.01` and `living_thing.n.01`. We retrieved a total of 14,942 HN and 17,912 non-HN.

For the LR and XGBoost models, we calculated four different scores to be used as features (see Appendix B for feature details and Appendix D for an ablation study). For the Transformer model, we used CamemBERT (Martin et al., 2020) and forwarded tokenized $\vec{W}$ as input.

We trained all three models on our golden HN and non-HN datasets with a 80/20 training-validation split using an NVIDIA RTX 4070 Super GPU. Hyperparameter optimization search was performed on all models using Weight & Biases (Biewald, 2020). Chosen hyperparameters can be found in Appendix C. An accuracy of 0.914 was achieved for the LR model; 0.937 for the XGBoost model; 0.927 for the Transformer model.

After classification, we retrieved 2,312 entries with 100% humanness agreement from Demonette, and 1,428 from TLFi recursive search results. Our final HN database contains 16,652 unique nouns across all six initial datasets (Wiktionary, Wikidata, NHUMA, Demonette, TLFi and TLFi Recursive).

3.2 Instruction/Output Datasets and Masculine-Specific Filtering

Our main objective is to comprehensively evaluate the extent to which LLMs are prone to using MG in their answers to generic instructions. To this goal, we leveraged the French HN database introduced in Section 3.1 to build a MG subset (5,140 entries), that is masculine-only HNs, specifically for instruction/output filtering purposes. This subset was created by removing epicene (i.e., words whose masculine and feminine forms are identical) and feminine words from the original HN database.

Then, to analyze the use of MG by LLMs, we resorted to four LLM instruction/output datasets shown in Table 1: oracle⁹; oasst2¹⁰; hh_rlh¹¹; and

⁹<https://github.com/opinionscience/InstructionFr/tree/main/wikipedia>

¹⁰<https://huggingface.co/datasets/OpenAssistant/oasst2>

¹¹https://huggingface.co/datasets/AIfll/french_hh_rlh

⁷This could happen because HNs found during the first search may also have a non-human meaning; for example “distributeur” can be translated both to “distributor” (human) and “dispenser” (non-human).

⁸Demonette counts exclude demonyms (21,340 entries).

Name	Type	Description	Pairs	Exp. 1 Filtered → Narrowed	Exp. 2 Filtered
oracle	Human	Set of question-answer pairs from French Wikipedia users	4,613	2,600 → 605	1,338
oasst2	Human	Open Assistant Conversations Dataset Release 2 French pairs	11,773	311 → 72	38
hh_rlhf	LLM	French translation of Anthropic’s hh-rlhf dataset, “chosen” pairs	161,000	10,806 → 2,520	1,515
alpaca	LLM	French instruction/output pairs generated by OpenAI’s GPT-3.5	55,184	29,179 → 6,803	1,665
Total			232,570	42,896 → 10,000	4,556

Table 1: Instruction/output datasets used for analyzing MG use. For both experiments, pairs were filtered to remove masculine specific uses (see Section 3.2). Pairs for Experiment 1 were further narrowed to 10,000 total to reduce LLM querying costs.

instruction	
A	<u>Mon docteur</u> m'a prescrit des somnifères, et... (My <u>doctor</u> [masc.] prescribed sleeping pills, and...)
B	<u>Qui est Albert Camus</u> ? (Who is <u>Albert Camus</u> ?)
output	
C	Il est vrai que le <u>duo</u> a fait une belle performance... (The <u>duo</u> indeed gave a great performance...)
D	Pour fixer un téléviseur au mur, vous devez d'abord... (To mount a TV to the wall, you first need to...)

Figure 3: Examples of filtered instructions/outputs. Example A features specific masculine due to the use of a possessive determiner. Example B is filtered due to name usage. Example C contains an ambiguous term.

alpaca¹². The first two are human-written instructions/outputs datasets, while the last two are LLM-generated/translated. Human outputs are included to gauge the extent to which French speakers use MG in written texts, and thus have a baseline when comparing with LLM outputs.

All datasets were filtered to remove entries exhibiting specific uses of the masculine gender (i.e., contexts where a masculine word does indeed refer to a male individual), which cannot be interpreted generically and are not MG. Examples of filtered instructions and outputs are shown in Figure 3, and examples of final generic instructions are made available in Appendix E. Filtering steps are further detailed below.

Using spaCy v3.8 (Honnibal et al., 2023) and the French spaCy Transformer model fr_dep_news_trf for state-of-the-art performance, we performed POS tagging and dependency

parsing on instructions from our four datasets in Table 1. We filtered instructions including a singular possessive determiner (e.g., “mon”, ‘my’) or a definite determiner (e.g., “ce”, ‘this’) followed by an HN (Ex. A in Figure 3). Similarly, instructions containing the interrogative pronoun “qui” (“who”), strictly used to refer to people, were removed as they could bias answers towards referring to a specific person; the same was applied to instructions and outputs with names (Ex. B).¹³

Finally, we specifically filtered a certain number of nouns found to be ambiguous during our preliminary tests¹⁴. For instance, the noun “duo” in Ex. C is used to refer to several known persons in context. The list was extended by scraping a part of Wiktionary’s list of 1,750 most common French words¹⁵ (1,229 entries in our list in total). Moreover, as we had previously used data from Wiktionary to fill our MG dictionary, we removed all nouns from Wiktionary whose human-related definition index was greater than 2¹⁶. This helped reduce the number of nouns which were not primarily and/or inherently considered or used as HNs.

¹³Only for the “oracle” dataset, we also left out parts of instructions with “oracle” or “pythie”, jargon used by the French Wikipedia community to refer to users answering inquiries.

¹⁴Full list of filtered nouns in the repository: https://github.com/spidersouris/llm_masc_gen.

¹⁵https://fr.wiktionary.org/wiki/Annexe:Liste_de_1750_mots_français_les_plus_courants

¹⁶According to the style guide of Wiktionary (https://en.wiktionary.org/wiki/Wiktionary:Style_guide#Definitions), the most common word definitions are placed first. By retrieving only the top 2 entries, we are making sure we are only using nouns commonly used to refer to human beings.

¹²<https://huggingface.co/datasets/jpacifico/French-Alpaca-dataset-Instruct-55K>

3.3 LLM Querying and HN Validation

To gauge the use of MG by LLMs, we query a set of LLMs of different sizes and types (both proprietary and local) with the filtered instructions. Then, we automatically detect and analyze the usage of MG and gender-fair (inclusive and neutral) terms in both LLM outputs and human-written outputs for comparison. As a post-processing step, we leverage GPT-4o mini¹⁷ and in-context learning (Brown et al., 2020) to validate nouns considered HNs in context. This was done because, during preliminary experiments, many nouns considered MG were in fact contextually ambiguous nouns not always referring to human beings.¹⁸

For HN validation, GPT-4o mini was chosen for its great performance/cost ratio. The words and the context in which they appear were forwarded to GPT-4o mini. We used Pydantic (Colvin et al., 2025) to constrain its output to JSON format, each noun being a key and its value being 0 if classified as non-human, or 1 if classified as human in context. We set max tokens to 500, temperature to 0. See Appendix K for prompting details. GPT-4o mini HN validation was evaluated by extracting 1,000 instructions from the hh_rlhf and alpaca datasets (see Section 3.2) and applying our MG use analysis pipeline. We narrow down the pipeline results to 500 to only retrieve contexts where at least one HN has been found. Those results are then sent to GPT-4o mini for validation. Two annotators classified the nouns in the 500 contexts as HN or non-HN (2,013 nouns in total). Cohen’s kappa (Cohen, 1960), noted κ , was used to compute inter-annotator agreement, where $\kappa = 0.944$. We built a reference annotation dataset and compared it with GPT-4o mini’s output, where $\kappa = 0.855$. Results indicate high agreement. Divergences with GPT-4o mini are mostly due to nouns used to refer to animal or anthropomorphic characters (examples in Table 19).

3.4 Measuring MG Bias

To take into account the fact that gender fairness can be expressed through multiple, partly competing linguistic strategies (inclusive markers, neutral markers and reduced use of MG), we model MG bias as a three-dimensional multiobjective opti-

mization problem, drawing on the theory of Pareto optimality (Pareto, 1906). Each objective is associated with a distinct linguistic strategy for mitigating MG bias:

$$\hat{M} = \frac{MG}{HN}, \quad \hat{I} = \frac{I}{HN}, \quad \hat{N} = \frac{N}{HN}$$

where MG is the total number of MG across all texts; I includes noun and pronoun inclusive forms (e.g., “il ou elle”, ‘he or she’; or “danseur-euses” to mean “male and female dancers”) as well as neopronouns such as “iel” or “celleux” (equivalents of “they” or “zir” in English); N the number of neutral nouns, which gender does not depend upon the referent’s (e.g., “personne”, ‘person’); and HN the total number of human nouns. Lower values of $\hat{M}$ indicate less reliance on MG, while higher values of $\hat{I}$ and $\hat{N}$ reflect greater use of inclusive and neutral GFL markers, respectively. These markers are detected using regular expressions (see Appendix L for details). All ratios are reported in percent.

We say that a model A Pareto-dominates a model B if $\hat{M}_A \leq \hat{M}_B$, $\hat{I}_A \geq \hat{I}_B$, $\hat{N}_A \geq \hat{N}_B$ with at least one strict inequality. A model is Pareto optimal (PO) if no other model in the comparison set dominates it. We also measure MG bias by computing MG use rate in responses and comparing with GFL markers use.

4 Experiments

We design two experiments. In Exp. 1, we prompt LLMs with instructions containing no MG (42,896 instructions narrowed to 10,000 dataset-stratified to reduce computing and API call costs), and we compare outputs with human-written, masculine-specific-filtered responses from datasets oasst2 and oracle. In Exp. 2, we prompt LLMs with instructions containing MG only (4,556 instructions). Count details are in Table 1. Filtering steps explained in Section 3.2 were applied to only keep generic outputs. Table in Appendix G shows the number of responses after filtering for both experiments. Then, HNs in responses were validated using GPT-4o mini (Section 3.3). Finally, all LLM outputs had their MG bias measured (Section 3.4). Steps were also applied to human-written outputs from datasets oasst2 and oracle for comparison.

For both experiments, 6 LLMs were evaluated (3 proprietary models: GPT-4o mini (OpenAI, 2024), Claude 3 Haiku (Anthropic, 2024) and Gemini 1.5 Flash (Gemini Team et al., 2024); 3 local: Llama

¹⁷<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

¹⁸For instance, in French, “facteur” can refer to a mailman, but can also mean “factor, cause”. Similarly, “navigateur” can refer either to a (male) sailor or to an Internet browser.

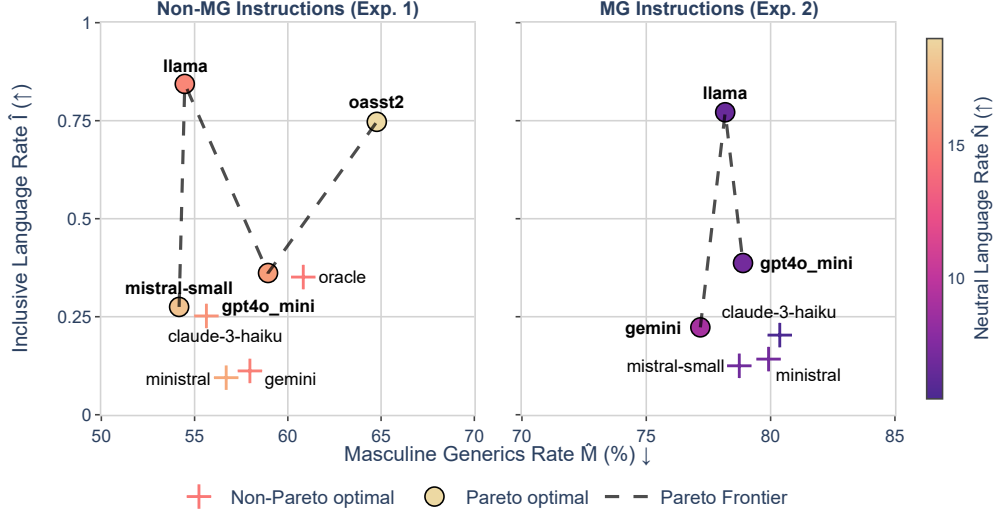


Figure 4: Pareto Frontier visualization for Experiments 1 and 2

Model/Dataset	$\hat{M} \downarrow$	$\hat{I} \uparrow$	$\hat{N} \uparrow$	PO?	P(PO) $\uparrow$
Experiment 1					
claude-3-haiku	55.63	0.25	15.69		0.38
gemini	57.97	0.11	14.86		0.01
gpt4o_mini	58.94	0.36	16.22	✓	0.55
llama	<u>54.47</u>	0.84	15.21	✓	1.00
ministral	56.69	0.09	16.74		0.23
mistral-small	54.17	0.28	<u>17.99</u>	✓	0.98
oasst2	64.77	<u>0.75</u>	19.02	✓	0.96
oracle	60.83	0.35	14.51		0.14
Experiment 2					
claude-3-haiku	80.37	0.20	5.45		0.00
gemini	77.17	0.22	8.91	✓	1.00
gpt4o_mini	78.89	<u>0.39</u>	<u>6.95</u>	✓	0.79
llama	<u>78.17</u>	0.77	6.56	✓	1.00
ministral	79.92	0.14	6.94		0.05
mistral-small	78.74	0.12	6.86		0.05

Table 2: Pareto results. Checked models are Pareto optimal. P(PO) is 95% certainty on 10,000 resamples. Best values are bolded, second-best are underlined.

3 8B Instruct (Touvron et al., 2023), Ministral 8B (Mistral AI, 2024) and Mistral Small 3 (Mistral AI, 2025)) using OpenRouter¹⁹. Models were chosen to evaluate an equal amount of proprietary and local models, and based on availability through the OpenRouter API and the cost/performance ratio.

5 Results

Table 2 reports the Pareto analysis results for both experiments, visualized in Figure 4. In Exp. 1 (non-MG prompts), 4 models are Pareto optimal: gpt4o_mini, llama, mistral-small, and oasst2. llama and oasst2 achieve the highest $\hat{I}$, while mistral-

small has the highest $\hat{N}$. Bootstrap analysis on 10,000 resamples shows high robustness for llama, mistral-small, and oasst2 ($P(\text{PO}) \geq 0.96$), whereas gpt4o_mini is on the Pareto frontier less consistently ($P(\text{PO}) = 0.55$). In Exp. 2 (MG prompts), 3 models are Pareto optimal: gemini, gpt4o_mini, and llama. llama displays once again the highest $\hat{I}$, while gemini achieves the lowest $\hat{M}$. All three models show high robustness under resampling, with $P(\text{PO})$ values ranging from 0.79 to 1.00. No single model dominates across all three dimensions in either experiment.

Figure 5 shows MG usage rates across models and datasets. On average, LLMs use MG in 27.57% of responses to non-MG prompts (Exp. 1), rising to 78.55% when prompts contain MG (Exp. 2). In Exp. 1, oasst2 shows significantly lower MG use rate than all LLMs based on non-overlapping confidence intervals (CIs), while differences among LLMs are not statistically reliable due to overlapping intervals. Restricting to responses with HNs only substantially increases MG rates to approximately 69-76% across all datasets, with overlapping CIs indicating no significant differences between models or human datasets in this condition. llama has the largest gap between overall MG rate and MG rate in responses with HNs (23.06 pp).

Figure 6 indicates that LLMs are reluctant to using inclusive GFL markers spontaneously (except llama, significantly). Neutral markers are a favored option, although we see no significant difference between models (except gemini in Exp. 2, significantly).

¹⁹<https://openrouter.ai/>

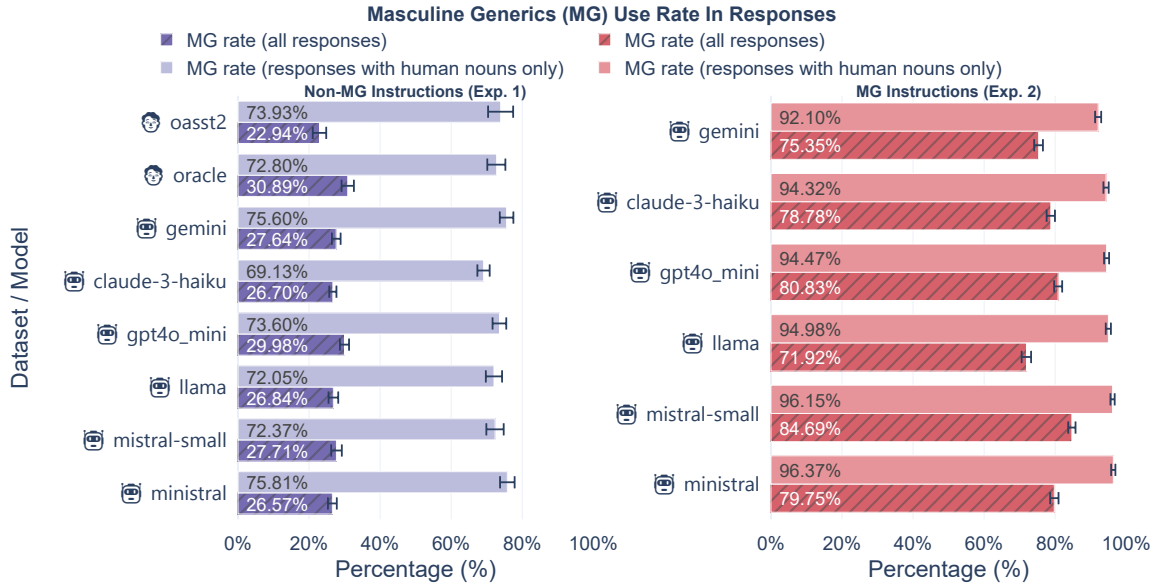


Figure 5: Percentage of MG use rate by dataset/model with $CI_{95\%}$ for Experiments 1 and 2. See Appendix F for CI details.

6 Discussion

Comparing LLM and human-written responses (Exp. 1) shows that human-written texts are not necessarily less MG-biased than LLM outputs. Indeed, while oasst2 is PO, it still has the highest $\hat{M}$ compared to all models, and oracle is not on the frontier. In contrast, llama appears as more gender-fair than human-written datasets and, more generally, as the most gender-fair LLM due to its high $\hat{I}$ and low $\hat{M}$, confirmed in both experiments.

Still, bias rate analysis shows that human-written dataset oasst2 has significantly lower overall MG bias than all LLMs. In contrast, many LLMs show overall bias rates comparable to or even lower than the oracle dataset, although not always significantly given overlapping CIs. The high MG use rate in oracle may come from the fact that the dataset contains answers to questions from Wikipedia users, mostly referring to content from published articles where contextual references may inflate MG counts. Thus, while all LLMs consistently exceed the least biased human baseline (oasst2), they perform comparably to oracle, indicating that LLM bias falls between the two human datasets rather than uniformly exceeding all human-written texts. Other, larger human datasets should be used for further comparison. Similarly, when considering responses with HNs only, most models exceed or match the rates of human datasets.

MG in prompts (Exp. 2) significantly increase

the prevalence of MG in LLM outputs, all while reducing GFL. While such an increase is expected for autoregressive models, it also shows the negative impact on other markers, and raises concerns about the downstream propagation of MG bias through human–LLM interaction when humans resort to MG. This highlights how users also need to be made aware of the consequences of using MG.

An analysis of GFL markers in both experiments highlights that inclusive markers are especially uncommon across all models and human-written datasets, with largely overlapping CIs (Table 12 in Appendix F). llama stands out due to its more frequent and statistically significant use of compact feminine endings²⁰, though a qualitative analysis on output examples shows that these are sometimes used incorrectly or applied to non-HN (Table 23 in Appendix L.1 for details). Regarding neopronouns, we find that no model uses them outside of explicit explanation contexts (e.g., claude-3-haiku in Table 24). This likely reflects their limited presence in training data, but it also means that LLMs do not naturally produce forms that go beyond the male–female binary. Neutral strategies, on the other hand are more common, mainly through the use of the metonymic noun “personne” (‘person’).

Additional analyses have been conducted on the

²⁰For example, for MG noun “vendeur” (‘seller’), the feminine suffix “euse” (from “vendeuse”) can be appended enclosed in parentheses in a compact form, as such: “vendeur(euse)”.

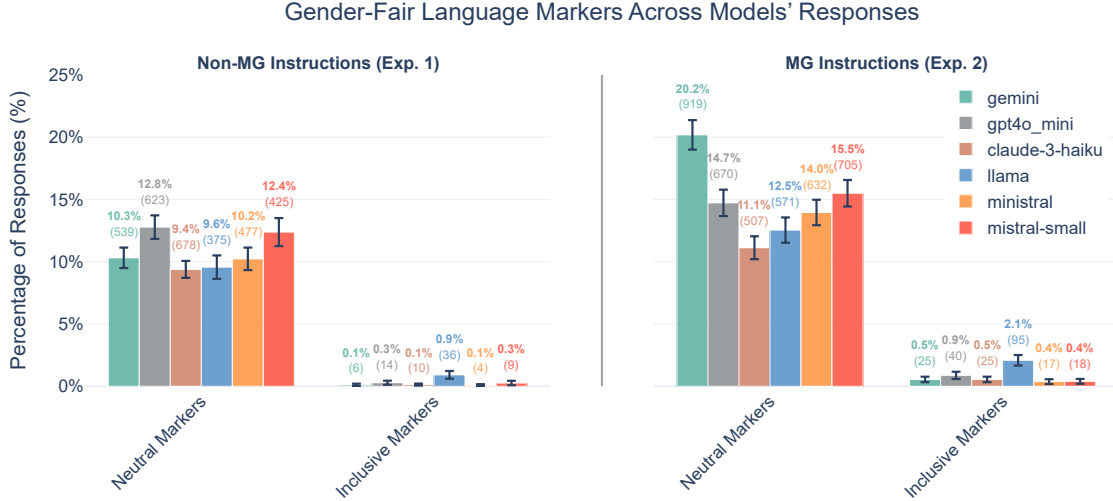


Figure 6: Percentage of responses with GFL markers across models with $CI_{95\%}$ for Experiments 1 and 2. See Appendix F for CI details.

types of HN used by LLMs in their responses, most particularly with regard to noun class (i.e., whether the noun refers to an occupation, a title...) and to usage specificity by LLMs. When it comes to noun class, most used HNs are occupation nouns (details in Appendix H). Relatedly, we also list the top 15 most common MG nouns used by all LLMs across both experiments in Appendix I. For usage specificity, we compute a noun-specific z -score to find outliers in MG usage by LLMs, and find out that some nouns are much more used by some LLMs, such as “entrepreneurs” by llama or “assistant” by ministral (details in Appendix J).

Overall, our results indicate that LLMs consistently exhibit MG bias when generating responses to generic instructions, whether MG are part of the prompt. While previous research has clearly demonstrated that LLMs display gender bias in specific or constrained contexts, this work provides evidence that MG-induced bias can be found in everyday, instruction-based interactions with LLMs. The prevalence of MG is concerning, as it reinforces existing gender bias, strengthens male mental representations and defies efforts towards inclusion and equality. These results show that fairness in language should be attentively considered when training LLMs in languages featuring MG.

Gender bias related to the use of MG may be reduced by rewriting MG-biased texts used for training through counterfactual data augmentation (Lu et al., 2020), a task known as “gender rewriting”. For example, Sun et al. (2021)

and Vanmassenhove et al. (2021) have created resources to rewrite masculine gendered pronouns in English to their gender-neutral equivalents (they/them). Similar approaches have been developed and extended to other parts of speech for Portuguese (Veloso et al., 2023), French (Doyen and Todirascu, 2025; Lerner and Grouin, 2024), and Italian (Greco et al., 2025; Piergentili et al., 2025). More recent approaches focus on using instruct-based LLMs for this task, leveraging techniques such as Retrieval-Augmented Generation (Muthusamy Chinnan et al., 2025) or task-specific instruction tuning (Wróblewska and Żuk, 2025).

7 Conclusion

Our work expands on prior research by analyzing gender bias in everyday, instruction-based interactions with LLMs. Using a French HN database created with HScorer and evaluating MG use by LLMs compared with human datasets, we find that LLMs consistently exhibit MG bias comparably to other human datasets. When evaluated in a multi-objective space, LLMs do not systematically dominate human data, showing that MG bias manifests differently across multiple dimensions. Notably, we notice a crucial lack of GFL inclusive markers use by LLMs; only Llama 3 8B Instruct marginally shows higher use of inclusive markers compared to other models. These findings highlight the need to control MG bias through multiple dimensions, and to promote GFL in LLM training data and/or during generation.

Limitations

Our work has several limitations. First, the French HN database that we created is not exhaustive and may have missing nouns. Similarly, even though we took care in not adding non-HNs to our database, a very small number of words not used to refer to human beings may be present in the data. Second, even though we took several steps to validate human nouns, both by pre-filtering our datasets and using an LLM for automatic verification (which we evaluated), many human nouns are polysemic, and some nouns may have incorrectly been detected as human nouns, or incorrectly left undetected. Third, detection of masculine specific uses remains a challenging task. Even though we designed several filtering steps to remove instructions and responses which may contain specific uses of the masculine so as to not bias the results, a small portion might still have been included to the data. Finally, our analysis only focuses on a small set of local and proprietary LLMs, and our comparison with human-written responses is limited by only 2 French datasets. More models and more human-written datasets should be analyzed to have a broader overview of MG bias in LLMs and a better comparison point with real usage.

Ethics Statement

The LLM instruction/output datasets were not filtered for harmful or malicious content. Similarly, LLM responses to generic instructions were not checked or filtered for such content. In addition, a very small portion of the HN database we created contains slurs and pejorative or discriminative terms used to refer to human beings. We deliberately did not filter these nouns, as they may be used in non-prejudicial contexts or with a non-harmful meaning. Finally, results of this work might be used in a counteractive way to increase gender bias in LLMs, for instance by removing occurrences of GFL markers from the training data or by increasing the number of MG occurrences in responses to generic instructions.

Acknowledgements

We would like to thank the reviewers for their insightful and valuable comments and feedback that greatly contributed in improving this paper.

This work of the Interdisciplinary Thematic Institute LIRIC, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm,

was supported by IdEx Unistra (ANR-10-IDEX-0002), and by SFRI-STRAT’US project (ANR-20-SFRI-0012) under the framework of the French Investments for the Future Program.

References

- Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.
- Lukas Biewald. 2020. Experiment tracking with weights and biases.
- Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Friederike Braun, Sabine Sczesny, and Dagmar Stahlberg. 2005. [Cognitive Effects of Masculine Generics in German: An Overview of Empirical Findings](#). *Communications*, 30(1):1–21.
- Creston Brooks, Samuel Eggert, and Denis Peskoff. 2024. [The rise of AI-generated content in Wikipedia](#). In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 67–79, Miami, Florida, USA. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Aaron Chatterji, Thomas Cunningham, David Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How People Use ChatGPT](#). Technical Report w34255, National Bureau of Economic Research, Cambridge, MA.
- Tianqi Chen and Carlos Guestrin. 2016. [XGBoost: A Scalable Tree Boosting System](#). In *Proceedings of the 22nd ACM SIGKDD International Conference*

- on *Knowledge Discovery and Data Mining*, pages 785–794.
- Jacob Cohen. 1960. [A Coefficient of Agreement for Nominal Scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, Alex Hall, and Victorien Plot. 2025. Pydantic.
- YiTian Ding, Jinman Zhao, Chen Jia, Yining Wang, Zifan Qian, Weizhe Chen, and Xingyu Yue. 2025. [Gender bias in large language models across multiple languages: A case study of ChatGPT](#). In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 552–579, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. [The Impact of AI-Generated Text on the Internet](#). *Preprint*, arXiv:2604.26965.
- Michael Döll, Markus Döhring, and Andreas Müller. 2024. [Evaluating Gender Bias in Large Language Models](#). *Preprint*, arXiv:2411.09826.
- Enzo Doyen and Amalia Todirascu. 2025. [GeNRe: A French Gender-Neutral Rewriting System Using Collective Nouns](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7889–7909, Vienna, Austria. Association for Computational Linguistics.
- Fanny Ducel, Aurélie Névéal, and Karën Fort. 2024a. Évaluation automatique des biais de genre dans des modèles de langue auto-régressifs. *TALN 2024*.
- Fanny Ducel, Aurélie Névéal, and Karën Fort. 2024b. La recherche sur les biais dans les modèles de langue est biaisée : état de l’art en abyme. *Revue TAL : traitement automatique des langues*, 64(3):119–143.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *Preprint*, arXiv:2403.05530.
- Sarah Glim, Anita Körner, and Ralf Rummer. 2024. [Generic masculine role nouns interfere with the neural processing of female referents: Evidence from the P600](#). *Language, Cognition and Neuroscience*, 39(10):1366–1375.
- Salvatore Greco, Moreno La Quatra, Luca Cagliero, and Tania Cerquitelli. 2025. [Towards AI-Assisted Inclusive Language Writing in Italian Formal Communications](#). *ACM Transactions on Intelligent Systems and Technology*, 16(4):1–24.
- Pascal Gygax, Ute Gabriel, Arik Lévy, Eva Pool, Marjorie Grivel, and Elena Pedrazzini. 2012. [The masculine form and its competing interpretations in French: When linking grammatically masculine role names to female referents is difficult](#). *Journal of Cognitive Psychology*, 24(4):395–408.
- Pascal Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. [Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men](#). *Language and Cognitive Processes*, 23(3):464–485.
- Pascal Gygax, Sayaka Sato, Anton Öttl, and Ute Gabriel. 2021. [The masculine form in grammatically gendered languages and its multiple interpretations: A challenge for our cognitive system](#). *Language Sciences*, 83:101328.
- Matthew Honnibal, Ines Montani, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Lisa Irmen and Julia Kurovskaja. 2010. [On the Semantic Content of Grammatical Gender and Its Impact on the Representation of Human Referents](#). *Experimental Psychology*, 57(5):367–375.
- Hadas Kotek, Rikker Dockum, and David Sun. 2023. [Gender bias and stereotypes in large language models](#). In *Proceedings of the ACM Collective Intelligence Conference, Ci ’23*, pages 12–24, New York, NY, USA. Association for Computing Machinery.
- Paul Lerner and Cyril Grouin. 2024. [INCLURE: A Dataset and Toolkit for Inclusive French Translation](#). In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 59–68, Torino, Italia. ELRA and ICCL.
- Jean-Daniel Lévy, Gaspard Lancrey-Javal, and Morgane Hauser. 2017. L’écriture inclusive : la population française connaît-elle l’écriture inclusive ? Quelle opinion en a-t-elle ? Rapport d’étude, Harris Interactive.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. [On LLMs-driven synthetic data generation, curation, and evaluation: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.
- Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. [Gender Bias in Neural Natural Language Processing](#). In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiko Okada, editors, *Logic, Language, and Security*, volume 12300, pages 189–202. Springer International Publishing, Cham.

- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonste de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [CamemBERT: A Tasty French Language Model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.
- Mistral AI. 2024. Un Ministral, des Ministraux. <https://mistral.ai/news/ministraux/>.
- Mistral AI. 2025. Mistral Small 3. <https://mistral.ai/news/mistral-small-3/>.
- Shunmuga Priya Muthusamy Chinnan, Meghann Drury-Grogan, and Bharathi Raja Chakravarthi. 2025. [Gender inclusive language generation framework: A reasoning approach with RAG and CoT](#). *Knowledge-Based Systems*, 328:114092.
- Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. [Synthetic Data Generation Using Large Language Models: Advances in Text and Code](#). *IEEE Access*, 13:134615–134633.
- Fiammetta Namer, Nabil Hathout, Dany Amiot, Lucie Barque, Olivier Bonami, Gilles Boyé, Basilio Calderone, Julie Cattini, Georgette Dal, Alexander Delaporte, Guillaume Duboisdindien, Achille Falaise, Natalia Grabar, Pauline Haas, Frédérique Henry, Mathilde Huguin, Nyoman Juniarta, Loïc Liégeois, Stéphanie Lignon, and 11 others. 2023. [Démonette-2, a derivational database for French with broad lexical coverage and fine-grained morphological descriptions](#). *Lexique*, 33.
- OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- Barbara Papasseudi. 2023. Classification de Noms d’Humains : Classification automatique des noms de fonctions, de métiers et de spécialistes. Master’s thesis, Université de Strasbourg.
- Vilfredo Pareto. 1906. *Manuale di economia politica*. Piccola Biblioteca Scientifica, Milan.
- Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. Gender-neutral rewriting in Italian: Models, approaches, and trade-offs. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 899–910, Cagliari, Italy. CEUR Workshop Proceedings.
- Princeton University. 2010. About WordNet.
- Célia Richy and Heather Burnett. 2021. [Démêler les effets des stéréotypes et le genre grammatical dans le biais masculin : une approche expérimentale](#). *GLAD!*, 10.
- Patrick Rothermund and Fritz Strack. 2024. [Reminding May Not Be Enough: Overcoming the Male Dominance of the Generic Masculine](#). *Journal of Language and Social Psychology*, 43(4):468–485.
- Elena Sofia Safina. 2024. [Effects of Grammatical Gender on Gender Inferences: Experimental Evidence From Italian Common Gender Nouns](#). *International Journal of Linguistics*, 16(3):60.
- Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. [A decade of gender bias in machine translation](#). *Patterns*, 6(6):101257.
- Jasmin Schmidt and Claudia Hess. 2026. [Tackling the Generic Masculine: Evaluating Gender Neutrality of German AI-Generated Texts](#). *International Conference on Gender Research*, 9(1):249–256.
- Jeanette Silveira. 1980. [Generic masculine words and thinking](#). *Women’s Studies International Quarterly*, 3(2-3):165–178.
- Nasim Sobhani and Sarah Delany. 2024. [Towards Fairer NLP Models: Handling Gender Bias In Classification Tasks](#). In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 167–178, Bangkok, Thailand. Association for Computational Linguistics.
- Dagmar Stahlberg, Sabine Sczesny, and Friederike Braun. 2001. [Name Your Favorite Musician: Effects of Masculine Generics and of their Alternatives in German](#). *Journal of Language and Social Psychology*, 20(4):464–469.
- Dejan Stosic and Véronique Lagae. 2013. La base de données NHUMA: avancées, problèmes et perspectives. In *Journées d’étude NHUMA 4*, Strasbourg, France.
- Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. [They, Them, Theirs: Rewriting with Gender-Neutral English](#). *arXiv preprint*.
- Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. [Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22975–23005, Vienna, Austria. Association for Computational Linguistics.
- Julia Tibblin, Joost Van De Weijer, Jonas Granfeldt, and Pascal Gygax. 2023. [There are more women in “joggeur-euses” than in “joggeurs”: On the effects of gender-fair forms on perceived gender ratios in French role nouns](#). *Journal of French Language Studies*, 33(1):28–51.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [LLaMA: Open and Efficient Foundation Language Models](#). <http://arxiv.org/abs/2302.13971>. *Preprint*, arXiv:2302.13971.

- Eva Vanmassenhove. 2024. [Gender Bias in Machine Translation and The Era of Large Language Models](#). *Preprint*, arXiv:2401.10016.
- Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. [NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.
- Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. [A rewriting approach for gender inclusivity in Portuguese](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759, Singapore. Association for Computational Linguistics.
- Yiluo Wei and Gareth Tyson. 2024. [Understanding the impact of AI-generated content on social media: The pixiv case](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, Mm '24, pages 6813–6822, New York, NY, USA. Association for Computing Machinery.
- Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2021. Biais de genre dans un système de traduction automatique neuronale : une étude préliminaire. *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*, 1:11–25.
- Alina Wróblewska and Bartosz Żuk. 2025. [Integrating gender inclusivity into large language models via instruction tuning](#). *Preprint*, arXiv:2508.18466.
- Hualin Xiao, Brent Strickland, and Sharon Peperkamp. 2023. [How Fair is Gender-Fair Language? Insights from Gender Ratio Estimations in French](#). *Journal of Language and Social Psychology*, 42(1):82–106.
- Tatu Ylonen. 2022. Wiktextextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association.
- Zhiwen You, HaeJin Lee, Shubhanshu Mishra, Sullam Jeoung, Apratim Mishra, Jinseok Kim, and Jana Diesner. 2024. [Beyond Binary Gender Labels: Revealing Gender Bias in LLMs through Gender-Neutral Name Predictions](#). In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 255–268, Bangkok, Thailand. Association for Computational Linguistics.

A Data Collection for Dictionary Building

For TLFi, we built a custom scraper using the Playwright Node.js library²¹ to retrieve words used to refer to human beings. We restricted the search to entries having one of their definitions starting with gender-neutral or gender-specific nouns or pronouns commonly used in dictionary entries related to human beings (see Table 3). We also performed recursive search to get nouns whose definition started with one of the previously scraped nouns.

For Wiktionary, we used Wikitextextract (Ylonen, 2022)²² and applied the same filtering rule with all but one noun (“individu”, ‘individual’)²³.

Following the preliminary work of Lerner and Grouin (2024), we also used Wikidata to retrieve masculine-feminine HN pairs with a SPARQL query (see Appendix A.1).

Table 3: Nouns and pronouns used to perform definition search

Noun	English	Websites
personne	person	TLFi, Wiktionary
individu	individual	TLFi
quelqu’un	someone	TLFi, Wiktionary
homme	man	TLFi, Wiktionary
femme	woman	TLFi, Wiktionary

A.1 Wikidata SPARQL Query Used for Human Noun Database

Listing 1: Wikidata SPARQL Query adapted from Lerner and Grouin (2024)

```

SELECT
  ?entity
  ?entityLabel
  ?entityDescription
  ?entity_male_label
  ?entity_female_label
WHERE
{
  ?entity wdt:P2521
    ?entity_female_label .
  OPTIONAL {
    ?entity wdt:P3321

```

```

    ?entity_male_label .
  FILTER (
    LANG(?entity_male_label)
      = "fr"
  ).
}
SERVICE wikibase:label {
  bd:serviceParam
    wikibase:language "fr" .
}
FILTER (
  LANG(?entity_female_label)
    = "fr"
).
OPTIONAL {
  ?entity schema:description
    ?entityDescription
  FILTER (
    LANG(?entityDescription)
      = "fr"
  ).
}
# Remove entries w/ description
# "titre de noblesse" as many
# dups with city names, e.g.
# "Duc de Gênes", "Duc de Teck"
FILTER (
  ?entityDescription
    != "titre de noblesse"@fr
  ||
  !BOUND(?entityDescription)
).
}

```

²¹<https://playwright.dev/>

²²<https://kaikki.org/>

²³This noun was removed from the filtering list as we had a certain number of false positives, as it can be used as a synonym of “species” and refer to living beings other than humans.

B HScorer Scoring Functions

This section details the scoring functions used as features $\vec{x}$ for the LR and XGBoost models, along with the FastText (Bojanowski et al., 2017) vector for word embedding representation. We calculate four different scores to be used to classify each word W : WordNet Hypernym Score (H), WordNet Definition Score (D), FastText Score (F) and Suffix Score (S). The resulting scores, together with the FastText representation $\vec{v}(W)$ of W , are concatenated to form the final feature vector $\vec{v}(W)$:

$$\vec{v}(W) = [H(W) \oplus D(W) \oplus F(W) \oplus S(W) \oplus v_W]$$

WordNet Hypernym Score. Let $H(W) = (h_s, n_s)$, where:

$$h_s = \frac{\sum_{p \in P} \sum_{y \in p} 1_h(y)}{|P|}$$

$$n_s = \frac{\sum_{p \in P} \sum_{y \in p} 1_n(y)}{|P|}$$

Where P is the set of hypernym paths for W , $1_h(y)$ is the indicator function for human hypernyms, and $1_n(y)$ is the indicator function for non-human hypernyms.

WordNet Definition Score. Let I_h be a set of human indicator words and I_n a set of non-human indicator words (see Table 4). These indicator words are manually defined English words related to human (“someone”, “person”, “who”) or non-human (“object”, “plant”, “chemical”) entities, and are used to search in WordNet definitions. Since WordNet definitions are universally in English, we use English words. Let $D(W) = (h_d, n_d)$ where:

$$h_d = \frac{\sum_{y \in Y} \sum_{i \in I_h} f(i, d(y))}{|Y|}$$

$$n_d = \frac{\sum_{y \in Y} \sum_{i \in I_n} f(i, d(y))}{|Y|}$$

Where Y is the set of synsets for W and $f(i, d)$ counts occurrences of indicator i in definition d .

FastText Score. We define a set of human prototypes P_h and non-human prototypes P_n (see Table 4). These prototype words are manually defined French words prototypically related to human (“personne” (*person*), “homme” (*man*), “femme” (*woman*)) or non-human (“objet” (*object*), “chose” (*item*), “machine” (*machine*)) entities, and are used

to compute semantic similarity. The prototype sets are more restrictive than the indicator sets. Let $F(W) = (h_f, n_f)$ where:

$$h_f = \frac{\sum_{p \in P_h} \cos(v_W, v_p)}{|P_h|}$$

$$n_f = \frac{\sum_{p \in P_n} \cos(v_W, v_p)}{|P_n|}$$

Where v_W is the 300-dimension FastText vector for word W and $\cos(v_a, v_b)$ is the cosine similarity between vectors a and b .

Suffix Score. Finally, the suffix score captures whether W ends with a human noun suffix. Specifically, $S(W) = 1$ if W ends with one of the suffixes $\hat{s} \in \hat{S}$ (see Table 5), and $S(W) = 0$ otherwise.

The four scores are then concatenated with the FastText vector v_W to obtain the final feature representation $\vec{v}(W)$ used by the LR and XGBoost models.

Human	Non-Human
Prototypes	
personne, femme, humain, gens	homme, objet, chose, machine, animal, plante, substance, outil, appareil, sentiment
Indicators	
someone, person, specialist, practitioner, employee, profession, whose, character, leader, participant, sibling, relative	somebody, people, professional, expert, worker, occupation, who, personality, member, follower, parent, child, relative, object, thing, item, device, tool, machine, equipment, apparatus, instrument, substance, material, element, chemical, plant, animal, organism, species, concept, idea, notion, theory, principle, place, location, area, region, structure, amount, quantity, measure, unit, state, condition, situation, circumstance

Table 4: Human and Non-Human Prototypes and Indicators

Masculine	Feminine
ant	ante
ent	ente
ateur	atrice
ien	ienne
aire	aire
eron	eronne
iste	iste
if	ive
and	ande
in	ine
Ø	esse
on	onne
teur	trice
eur	euse
ier	ière
ais	aise
ois	oise
ard	arde
ain	aine
euf	euve
urge	urge
graphe	graphe
naute	naute
anthrope	anthrope
crate	crate
mane	mane
pathe	pathe
phile	phile

Table 5: Masculine and feminine suffixes used to compute S .

XGBoost

```

booster = gbtree
learning_rate = 0.22394632872649503
max_depth = 10
min_child_weight = 78
subsample = 1
colsample_bytree = 1
n_estimators = 912
gamma = 0
reg_alpha = 0
reg_lambda = 0
objective = binary:logistic
early_stopping_rounds = 20
random_state = 42
tree_method = gpu_hist
n_jobs = 24

```

Transformer

```

train_batch_size = 16
eval_batch_size = 16
eval_strategy = epoch
save_strategy = epoch
num_train_epochs = 5
weight_decay = 0.061748150962771656
learning_rate = 8.497821083760116 × 10-6

```

C HScorer Cross-Validation and Final Models Hyperparameters

Logistic Regression (LR)

```

penalty = l1
solver = saga
C = 100

```

D HScorer Feature Ablation Study

To gauge what are the most important features to the LR and XGBoost HScorer models, we conducted an ablation study. Table 7 shows our different experiments, testing configurations without specific features and other configurations where only one feature is used and computing precision (P), recall (R), accuracy (acc) and F-score ($F1$). Features (scores) are the same as specified in Appendix B. Table 8 also shows the drop in accuracy caused by the removal of specific features.

E Examples of Generic Instructions

Examples of generic instructions from the different datasets are shown in Table 9.

Table 6: HScorer 5-Fold Cross-Validation Results on Classic Machine Learning Models

Model	Precision	Recall	Accuracy	F1
CART	83.50 $\pm$ 0.42	83.66 $\pm$ 0.49	85.05 $\pm$ 0.27	83.58 $\pm$ 0.30
KNN	85.65 $\pm$ 0.49	88.09 $\pm$ 0.56	87.87 $\pm$ 0.43	86.85 $\pm$ 0.47
RF	93.98 $\pm$ 0.37	85.61 $\pm$ 0.49	90.96 $\pm$ 0.32	89.60 $\pm$ 0.38
LR	89.07 $\pm$ 0.36	91.23 $\pm$ 0.35	90.92 $\pm$ 0.19	90.14 $\pm$ 0.20
XGBoost	93.96 $\pm$ 0.39	91.85 $\pm$ 0.29	93.61 $\pm$ 0.22	92.89 $\pm$ 0.24

F Language Markers and MG Bias Rates Confidence Intervals

To quantify statistical uncertainty for estimated MG bias rates as well as for language markers by category, we estimated 95% bootstrap CIs using 10,000 resamples for each model and experiment.

MG bias rates. Estimates are shown in Tables 10 (Exp. 1) and 11 (Exp. 2). While all models/datasets exhibit substantial MG bias in both experiments, CIs show that, in Exp. 1, oasst2 has significantly lower overall bias than all LLM models. Among LLMs, however, overlapping CIs indicate that differences are not statistically significant, particularly in the middle of the ranking. In Exp. 2, llama demonstrates significantly lower bias

than the most biased models (mistral-small and gpt4o_mini), while the remaining models show overlapping intervals. For bias with human nouns (HN), CIs largely overlap across all datasets and models in both experiments, suggesting similar MG use with human referents.

Language markers. Estimates are shown in Table 12. For inclusive language markers in Experiment 1, all models show very low rates (below 1%). llama shows the highest rate at 0.92%, with a CI of [0.64, 1.23], making it distinguishable from most other models due to no overlap. However, for other models, the CIs overlap substantially due to very few occurrences. As a result, we cannot reliably say whether differences in these models’ inclusive markers rates reflect true differences or simply sam-

Table 7: HScorer Feature Ablation Study Results (LR and XGBoost Models)

experiment	P_{LR}	R_{LR}	Acc_{LR}	$F1_{LR}$	P_{XGB}	R_{XGB}	Acc_{XGB}	$F1_{XGB}$
all	0.903	0.906	0.913	0.904	0.936	0.924	0.937	0.930
wo_H	0.893	0.891	0.901	0.892	0.920	0.903	0.920	0.911
wo_D	0.901	0.903	0.911	0.902	0.933	0.922	0.934	0.928
wo_F	0.899	0.908	0.912	0.904	0.937	0.924	0.937	0.931
wo_S	0.899	0.908	0.911	0.903	0.935	0.925	0.936	0.930
wo_W	0.792	0.835	0.825	0.813	0.794	0.826	0.823	0.810
only_H	0.597	0.971	0.688	0.740	0.600	0.967	0.690	0.740
only_D	0.584	0.071	0.553	0.126	0.529	0.962	0.591	0.682
only_F	0.746	0.733	0.764	0.740	0.739	0.732	0.760	0.736
only_S	0.660	0.379	0.628	0.482	0.660	0.379	0.628	0.482
only_W	0.883	0.889	0.896	0.886	0.916	0.893	0.914	0.904

Table 8: HScorer Feature Importance (LR and XGBoost Models)

Feature	LR Acc. Drop	LR Drop %	XGB Acc. Drop	XGB Drop %
H	0.011262	1.233950	0.017045	1.819659
D	0.001978	0.216775	0.002435	0.259951
F	0.000913	0.100050	-0.000609	-0.064988
S	0.001217	0.133400	0.000304	0.032494
W	0.087962	9.638152	0.113681	12.136474

pling variation. The same applies to Experiment 2, where llama once again shows a statistically significant rate at 2.09% with a CI of [1.70, 2.51], much higher than other models.

When it comes to neutral language markers, in Experiment 1, gpt4o_mini [11.86, 13.73] and mistral-small [11.32, 13.51] are distinguishable from the lower cluster. In Experiment 2, gemini

Instruction	English Translation	Source
Comment certains médicaments provoquent-ils une prise de poids ?	How do some medications cause weight gain?	hh_rlh
Rédige une publicité de 50 mots pour un nouveau parfum de luxe.	Write a 50-word advertisement for a new luxury perfume.	alpaca
Les poissons rouges peuvent-ils survivre dans du sirop de menthe ?	Can goldfish survive in mint syrup?	oracle
Explique-moi les ordinateurs quantiques comme si j'avais 10 ans.	Explain quantum computers to me as if I were 10 years old.	oasst2
Quelles sont les conditions pour devenir sauveteur ?	What are the requirements to become a lifeguard ?	hh_rlh
Explique le rôle du président des États-Unis .	Explain the role of the President of the United States .	alpaca
Bonjour, je souhaiterais connaître la différence entre un décorateur d'intérieur et un architecte d'intérieur . Merci d'avance.	Hello, I would like to know the difference between an interior decorator and an interior designer . Thank you in advance.	oracle
Quels sont les stéréotypes les plus répandus concernant les français ?	What are the most common stereotypes about French people ?	oasst2

Table 9: Examples of generic instructions used to query LLMs from all four datasets. First 4 examples are non-MG prompts (Exp. 1); last 4 examples are MG prompts (Exp. 2) with MG nouns in bold.

Table 10: Bias rate estimates (in %) with 95% bootstrap CIs (Exp. 1)

Dataset	Overall Bias (%)	CI _{95%}	Bias with HN (%)	CI _{95%}
oracle	30.890	[29.150, 32.622]	72.800	[70.167, 75.259]
gpt4o_mini	29.980	[28.689, 31.254]	73.600	[71.646, 75.509]
mistral-small	27.710	[26.224, 29.225]	72.370	[69.901, 74.783]
gemini	27.640	[26.436, 28.886]	75.600	[73.681, 77.510]
llama	26.840	[25.485, 28.214]	72.050	[69.756, 74.323]
claude-3-haiku	26.700	[25.689, 27.711]	69.130	[67.379, 70.841]
ministral	26.570	[25.322, 27.811]	75.810	[73.756, 77.864]
oasst2	22.940	[21.045, 24.834]	73.930	[70.388, 77.446]

Table 11: Bias rate estimates (in %) with 95% bootstrap CIs (Exp. 2)

Model	Overall Bias (%)	CI _{95%}	Bias with HN (%)	CI _{95%}
mistral-small	84.690	[83.660, 85.727]	96.150	[95.551, 96.746]
gpt4o_mini	80.830	[79.705, 81.970]	94.470	[93.758, 95.172]
ministral	79.750	[78.578, 80.941]	96.370	[95.750, 96.946]
claude-3-haiku	78.780	[77.590, 79.939]	94.320	[93.576, 95.056]
gemini	75.350	[74.072, 76.555]	92.100	[91.231, 92.948]
llama	71.920	[70.562, 73.221]	94.980	[94.236, 95.686]

shows a statistically significant rate at 20.20% with a CI of [19.03, 21.38]. Otherwise, similarly to inclusive markers, other models show much overlap.

G Number of Non-Specific LLM Responses

Table 14: Number of LLM responses without specific masculine markers by model

Model	Experiments	
	Experiment 1	Experiment 2
claude-3-haiku	7,221	4,556
gemini	5,224	4,551
gpt4o_mini	4,873	4,548
ministral	4,660	4,528
llama	3,920	4,452
mistral-small	3,432	4,547

Table 12: Language markers estimates (in %) by category with 95% bootstrap CIs (Exp. 1)

Dataset	Incl. (%)	CI _{95%}	Neut. (%)	CI _{95%}
oracle	0.33	[0.14, 0.54]	9.62	[8.54, 10.74]
oasst2	0.31	[0.10, 0.56]	9.06	[7.78, 10.34]
gemini	0.11	[0.04, 0.21]	10.32	[9.51, 11.14]
gpt4o_mini	0.29	[0.14, 0.45]	12.79	[11.86, 13.73]
claude-3-haiku	0.14	[0.06, 0.22]	9.39	[8.72, 10.07]
llama	0.92	[0.64, 1.23]	9.57	[8.65, 10.51]
ministral	0.09	[0.02, 0.17]	10.24	[9.36, 11.14]
mistral-small	0.26	[0.12, 0.44]	12.39	[11.32, 13.51]

Table 13: Language markers estimates (in %) by category with 95% bootstrap CIs (Exp. 2)

Dataset	Incl. (%)	CI _{95%}	Neut. (%)	CI _{95%}
gemini	0.55	[0.35, 0.77]	20.20	[19.03, 21.38]
gpt4o_mini	0.88	[0.62, 1.17]	14.74	[13.72, 15.79]
claude-3-haiku	0.55	[0.35, 0.77]	11.13	[10.23, 12.05]
llama	2.09	[1.70, 2.51]	12.59	[11.66, 13.56]
ministral	0.38	[0.20, 0.57]	13.96	[12.97, 14.98]
mistral-small	0.40	[0.22, 0.59]	15.51	[14.47, 16.57]

H MG Human Noun Classes

We measure what types of HNs are most used as MG. To that end, we leverage the NHUMA dataset (used to create our HN dictionary in Section 3.1), which contains human-annotated HN types (789 annotated nouns with labels such as “profession”, “doer”, “relationship” or “status”; see Table 16 for the list of all labels). We complete these annotations using a spaCy model trained by Papasseudi (2023) on NHUMA data for automatic HN type labelling (see Table 15 for details) and we apply it to nouns in our MG dictionary missing human annotation (i.e., non-NHUMA entries). Entries not detected as HNs by the model were left unannotated. In total, our HN dataset contains 2,893 annotated entries (i.e., +2,104 after applying the model).

Table 15: Human noun classes and counts used by Papasseudi (2023) for spaCy model training

Class	Count	Mapped to
NH-Mét	697	profession
NH-Fonc	402	profession
NH-Spé	385	speciality
NH-Titre	111	title
NH-Grade	16	title

Table 16: Human noun classes and counts across entire human noun database (including non-MG)

Class	Count
profession	2,893
demonym	1,290
doer	388
speciality	269
attribute	179
relationship	44
status	24
title	18
patient	14
other	7
recipient	5

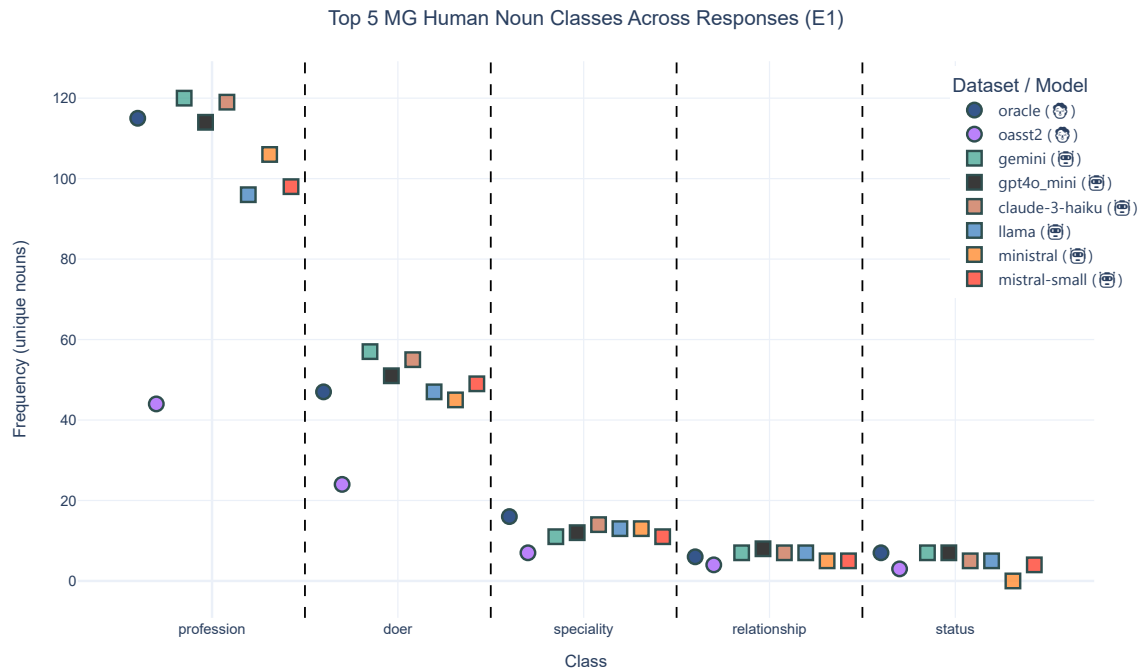


Figure 7: Frequency of unique MG nouns in responses based on their associated human noun class (Exp. 1)

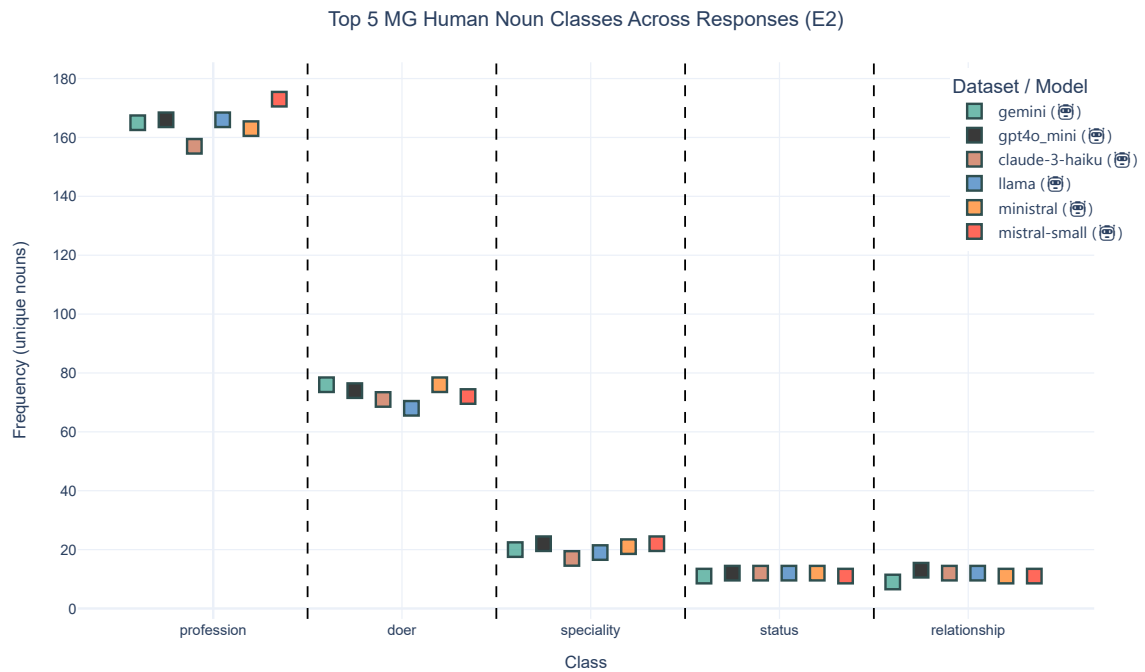


Figure 8: Frequency of unique MG nouns in responses based on their associated human noun class (Exp. 2)

I MG Counts

Table 17: Most used MG across LLM responses (Exp. 1)

Rank	Noun	English	Count
1	professionnel	professional	1,251
2	utilisateur	user	1,057
3	ami	friend	991
4	habitant	resident	977
5	joueur	player	767
6	client	customer	590
7	consommateur	consumer	535
8	citoyen	citizen	469
9	homme	man	409
10	humain	human	380
11	auteur	author	347
12	parent	parent	347
13	employé	employee	318
14	médecin	doctor	301
15	patient	patient	301

Table 18: Most used MG across LLM responses (Exp. 2)

Rank	Noun	English	Count
1	ami	friend	6,497
2	homme	man	3,589
3	étudiant	student	2,665
4	client	customer	2,452
5	invité	guest	2,391
6	utilisateur	user	2,159
7	auteur	author	2,021
8	président	president	1,988
9	monsieur	mister	1,667
10	joueur	player	1,629
11	adolescent	teenager	1,559
12	humain	human	1,533
13	parent	parent	1,242
14	acteur	actor	1,239
15	candidat	candidate	1,148

J MG Outliers with z -score

To determine which nouns are disproportionately frequent for certain LLMs, we identify outlier MG counts using a noun-specific z -score. The standard formula for z -score is:

$$z = \frac{x - \mu}{\sigma}$$

Where x is the observed value, μ the mean of all observations, and σ the standard deviation.

For each noun, we first compute the mean MG count across all observations:

$$\mu_{\text{noun}} = \frac{1}{n} \sum_{i=1}^n x_i$$

Where x_i denotes the MG count of the noun in observation i .

We then calculate the standard deviation of the MG counts for each noun:

$$\sigma_{\text{noun}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_{\text{noun}})^2}$$

The z -score for each observation is subsequently computed as:

$$z_i = \frac{x_i - \mu_{\text{noun}}}{\sigma_{\text{noun}}}$$

Nouns are classified as outliers when their z -score exceeds 1.7. This value was adjusted empirically. Results are displayed in Figure 9 for Experiment 1, and Figure 10 for Experiment 2.

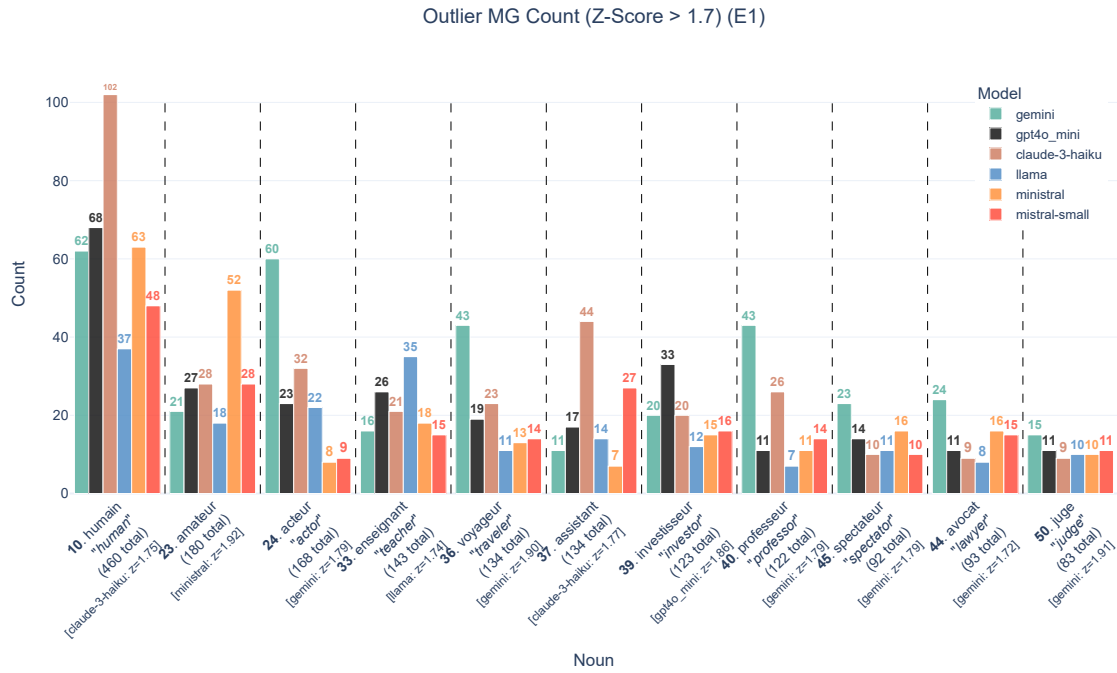


Figure 9: Counts of MG featuring model-specific outlying uses, with a z -score greater than 1.7 (Exp. 1)

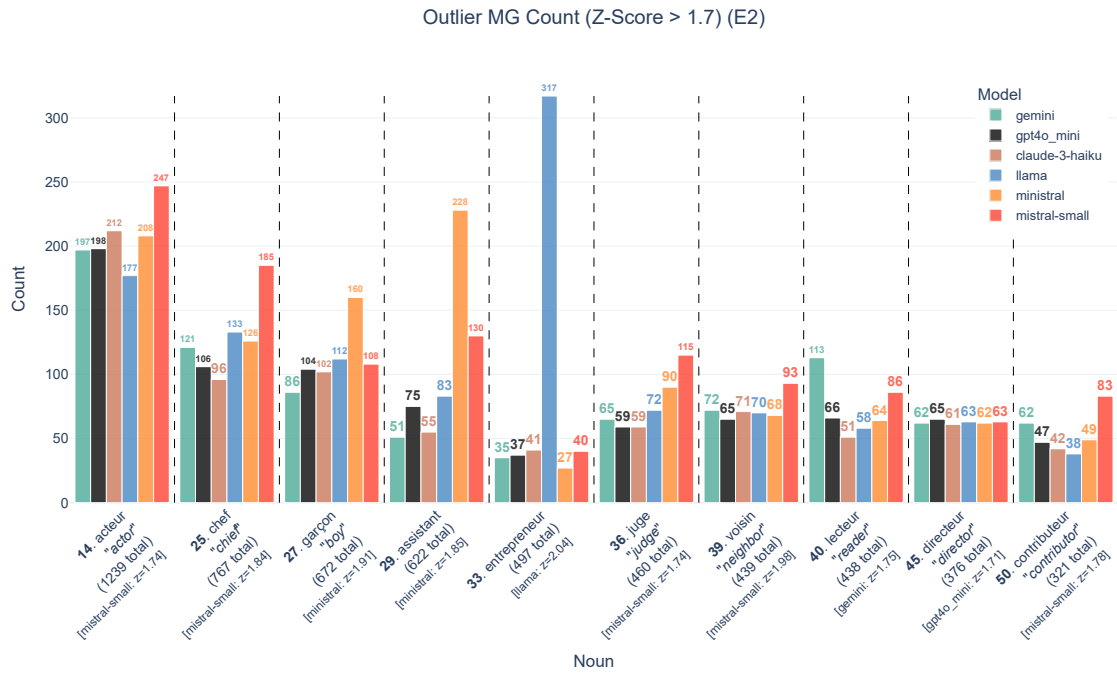


Figure 10: Counts of MG featuring model-specific outlying uses, with a z -score greater than 1.7 (Exp. 2)

K GPT-4o mini Human Noun Validation Prompting Details

Listing 2: System Prompt

You are an assistant that validates human noun classifications in French texts.

Listing 3: User Prompt

Given a text and nouns, for each noun, determine if it is a human noun in context. Some nouns may appear multiple times in the text. In such cases, they are distinguished by ID ('noun_1', 'noun_2'...), following the order in which they appear. Do not assume that all occurrences of the same noun are either human or non-human; instead, assess each occurrence individually based on its unique context. Only respond in this format, where human_noun is the noun being considered.

```
{{  
  "human_noun": 0,  
  "human_noun_2": 1  
}}
```

Examples

Text: Les facteurs d'employabilité des facteurs, chargés de distribuer le courrier, vont évoluer.

Nouns: facteurs, facteurs_2

Output: {{ "facteurs": 0, "facteurs_2": 1 }}

Text: Le président a annoncé aux citoyens une série de mesures pour renforcer l'économie du pays.

Nouns: président, citoyens, mesures

Output: {{ "président": 1, "citoyens": 1, "mesures": 0 }}

Text: Il croit aux esprits et aux fantômes depuis qu'il est enfant.

Nouns: esprits, fantômes, enfant

Output: {{ "esprits": 0, "fantômes": 0, "enfant": 1 }}

Text: {text}

Nouns: {human_nouns}

Output:

K.1 Examples of Disagreements with GPT-4o mini

Table 19 displays examples of disagreements between human annotators and GPT-4o mini on the 500 contexts sampled for validation. HNs incorrectly classified as such by GPT-4o are **in bold**.

Table 19: Examples of disagreements between human annotators and GPT-4o mini on HN annotation

Text	English Translation
Le papillon se repose un moment pour que ses ailes se développent et sèchent, avant de pouvoir s’envoler à la recherche de nourriture et de partenaires .	The butterfly rests for a while so that its wings can develop and dry before it can take flight in search of food and mates .
En chemin, ils rencontrèrent de nombreux alliés : un sage hibou qui leur enseigna des sorts puissants, un groupe de nains courageux qui leur forgèrent des armes magiques, et un vieux dragon sage qui leur révéla la vérité sur l’ombre maléfique.	Along the way, they encountered many allies : a wise owl who taught them powerful spells, a group of courageous dwarves who forged magical weapons for them, and an old, wise dragon who revealed to them the truth about the malevolent shadow.
Il est vrai que de nombreux films anciens de Disney montrent des personnages fumant, ce qui peut sembler étrange aujourd’hui, étant donné les connaissances sanitaires actuelles concernant le tabagisme.	It is true that many older Disney films depict characters smoking, which may seem strange today, given current health knowledge about smoking.

L GFL Markers

Table 20 displays examples of GFL markers automatically detected for the analyses. `fem_ending` refers to occurrences where the feminine ending of an HN is separated from the original, masculine form to increase its visibility. This is done either by adding a special separator (e.g., “auteur-ice”, “auteur(ice)”; see Tables 21 and 22) or by spelling the feminine ending with capitals (e.g., “auteurICE”).

We also include, in Table 21, the list of marker types included in the computation of each metric for the Pareto Frontier (Section 3.4). The corresponding separators used to detect the `fem_ending` type are listed in Table 22.

Example	English Translation	Type
mesdames et messieurs	Ladies and gentlemen	<code>incl_greetings</code>
tous et toutes	everyone	<code>incl_greetings</code>
un ou une	a [m] or a [f]	<code>incl_pairs</code>
il ou elle	he or she	<code>incl_pairs</code>
iel	they (sg.)	<code>neutral_prons</code>
auteur-ice; auteur(ice)	author	<code>fem_ending</code>

Table 20: Examples of inclusive markers by type, with their English equivalents

Type	Metric	Matching pattern	Examples
incl_greetings	<i>I</i>	Fixed inclusive greeting expressions, plural form	<i>tous et toutes, toutes et tous, messieurs, dames, messieurs dames, mesdames et messieurs, messieurs et mesdames</i>
incl_pairs	<i>I</i>	[masculine/feminine form] (et/ou) [corresponding form]	<i>ceux et celles, celles ou ceux, ils et elles, il ou elle, le ou la, un ou une</i>
fem_ending	<i>I</i>	Inclusive language separators (see Table 21)	<i>auteur·ices, auteur(ice)s ...</i>
neutral_prons	<i>I</i>	Gender-neutral pronouns	<i>iel, iels, ielle, ielles, ceux, elleux, ille, illes</i>
neutral_nouns	<i>N</i>	Common neutral nouns	<i>gens, personne, personnes, individu, individus</i>

Table 21: List of markers by type, with corresponding metric, matching pattern and examples

Glyph	Code point	Unicode Name
.	U+00B7	MIDDLE DOT
·	U+2027	HYPHENATION POINT
.	U+22C5	DOT OPERATOR
●	U+2219	BULLET OPERATOR
•	U+2022	BULLET
.	U+002E	FULL STOP
/	U+002F	SOLIDUS
(	U+0028	LEFT PARENTHESIS

Table 22: List of French inclusive language separators (placed between the stem and the feminine suffix, as in étudiant·es) for the fem_ending type

L.1 Examples of Gender-Fair LLM Outputs

We conduct a qualitative analysis of some LLM output examples from both experiments to determine how well they can use GFL in unconstrained and unprompted/open-ended contexts. Examples of incorrect GFL usage are listed in Table 23: the wrong GFL marker is bolded and the Table includes corresponding explanations as to why these usages are wrong. Errors mostly come from an overapplication of GFL techniques to nouns that do not need it (e.g., nouns referring to objects or non-human entities), or the use of a separator suffix that does not work because the feminine equivalent of the masculine noun is morphologically very different. Conversely, Table 24 shows examples of correct GFL usage by LLMs, with the GFL marker in bold and the corresponding passage English translation as a footnote.

Table 23: Examples of incorrect gender-fair LLM outputs by model and marker type

Model (Exp.)	Marker Type	Output	Explanation
gemini (E2)	incl_pairs	**Finaliser le lieu et la date avec les complices:** Vérifier avec le gérant du lieu ou l'hôte(ile).	“hôtel” (hotel) is an inanimate noun and does not change with gender, so using a gender separator is not appropriate. However, interestingly, the MG noun “gérant” (manager) used previously does not have the same separator applied, while it could have: “gérant(e)”.
llama (E2)	incl_pairs	Cher/e Monsieur(e) [Enseignant], Je vous écris pour solliciter une réunion avec vous concernant ma note obtenue	The model was asked to write a letter to a teacher. For the salutation, even though “Dear” (“cher”, “chère”) was correctly written separating the feminine suffix, the equivalent of “Monsieur” (“Mister”) for a woman is “Madame”, not “Monsieure”. Applying a separator that way does not work.
llama (E2)	incl_pairs	Bonjour! Je serais ravi de vous fournir une liste de vocabulaire français de base pour les débutants. Voici quelques notions essentielles pour commencer à apprendre le français [...] Soeur (sister) * Mari(e) (husband/wife)	The model was asked to provide a list of basic French vocabulary for a language learner. The equivalent of “mari” (“husband”) for a woman is “femme”, not “marie”. Applying a separator that way does not work.

Table 24: Examples of correct gender-fair LLM outputs by model and marker type

Model (Experiment)	Marker Type	Output
mistral-small (E1)	incl_pairs	Vous soulevez un point intéressant. L’expression “c’est une égalité parfaite” est souvent utilisée dans le langage courant pour renforcer l’idée de parité ou d’équilibre absolu entre deux choses. [...] Par exemple, si quelqu’un dit “c’est une égalité parfaite entre nos deux équipes,” il ou elle [*] veut dire que les deux équipes sont très équilibrées en termes de force et de capacité. Pour gérer ces situations, vous pouvez soulager ce sentiment en disant que “l’expression ‘de parfaite équité’ ou ‘d’équilibre parfait’ pourrait mieux exprimer l’idée d’une égalité absolue.”
gemini (E1)	fem_ending	Les médias sociaux, bien que présentant des avantages en termes de connexion et d’information, engendrent également des effets négatifs. La comparaison constante avec des versions idéalisées de la vie des autres utilisateur·ices [†] peut entraîner une baisse de l’estime de soi et de l’anxiété. Une exposition excessive à des contenus négatifs, polarisants ou manipulateurs peut affecter le bien-être mental et contribuer à la propagation de fausses informations. Enfin, la dépendance aux notifications et la distraction constante peuvent entraver la concentration et la vie sociale réelle.
llama (E1)	incl_greetings	Mesdames et messieurs [‡] , Aujourd’hui, je voudrais vous parler d’un sujet qui concerne tous les habitants de notre planète : la sauvegarde des espèces menacées. C’est un problème immense et urgent qui nécessite notre attention et notre action. Mais avant de vous présenter les raisons pour lesquelles la sauvegarde des espèces menacées est une question essentielle, je voudrais poser une question simple : qu’advierait-il si nous perdions définitivement certaines espèces qui vivent sur terre? [...]
claude-3-haiku (E2)	neutral_pron	Je comprends votre point de vue. La question de l’utilisation des pronoms peut être un sujet délicat et controversé. Voici quelques points à considérer : - L’utilisation des pronoms personnels (il/elle, iel [§] , etc.) est importante pour permettre aux personnes de s’identifier de la manière qui leur convient le mieux. [...]

^{*} English Translation: “For example, if someone says ‘there is perfect equality between our two teams,’ **he or she** means...”

[†] English Translation: “Continuously comparing one’s life with idealized versions of other **users**’ [**m./f.**] lives...”

[‡] English Translation: “**Ladies and gentlemen**...”

[§] English Translation: “The use of personal pronouns (he/she, **zir [equiv.]**, etc.) is important to help people identify themselves in the way that suits them best.”